DOCUMENT RESUME

ED 324 352 TM 015 609

AUTHOR Forsyth, Robert A.
TITLE The NAEP Proficiency Scales: Do They Yield Valid Criterion-Referenced Interpretations? Iowa Testing Programs Occasional Papers Number 35, May 1990.
INSTITUTION Iowa Testing Programs, Iowa City.
PUB DATE May 90
NOTE 29p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; *Criterion Referenced Tests; Elementary Secondary Education; Item Response Theory; *National Programs; *Rating Scales; Testing Problems; Testing Programs; *Test Interpretation; Test Validity
IDENTIFIERS *National Assessment of Educational Progress; Science Proficiency Scale

ABSTRACT

The validity of criterion-referenced interpretations of the proficiency scales of the National Assessment of Educational Progress (NAEP) is discussed. A major goal of the NAEP scales is to describe student achievement in specific content areas from grade 3 (age 9 years) through grade 11 (age 17 years). The numerical values for NAEP scales are intended to be interpreted almost exclusively from a criterion-referenced perspective. The NAEP intends to accomplish the goal of making criterion-referenced interpretations of examinee scores through the use of item response theory methodology. The basis for classifying an item at a proficiency level is the estimated probability of success of examinees at different proficiency levels. The implications for criterion-referenced interpretations are discussed in the context of the Science Proficiency Scale (SPS). Problems with interpreting SPS results are due to the following factors: (1) the science domain is not well defined; (2) it is doubtful that learning in this domain proceeds through an orderly sequence; (3) the number of test items that could be developed is virtually unlimited; and (4) many problems are generated by the interactions between the examinee's past experiences and item content. The NAEP mathematics, reading, and history scales are also reviewed. It is suggested that the NAEP proficiency scales do not yield valid criterion-referenced interpretations. Four figures illustrate the discussion. A 27-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert A. Forsyth
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED324352

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 35—May 1990

The NAEP Proficiency Scales: Do they yield valid criterion-referenced interpretations?

Robert A. Forsyth

TM015609

THE NAEP PROFICIENCY SCALES: DO THEY YIELD VALID CRITERION-REFERENCED INTERPRETATIONS?[1]

It seems reasonable to assert that no assessment program in our history has received the amount of public attention given to the National Assessment of Educational Progress (NAEP) in the last five years and to assert, further, that this attention will very likely increase during the next five years. It appears that NAEP data have recently been and will continue to be the major information base for making judgments about certain aspects of our educational system and for supporting selected policy decisions. Consider, for example, the three statements below, all of which were based on information presented in the latest NAEP *Science Report Card* (Mullis and Jenkins, 1988):

> More than half of the nation's 17-year-olds appear to be inadequately prepared either to perform competently jobs that require technical skills or to benefit substantially from specialized on-the-job training. The thinking skills and science knowledge possessed by these high-school students also seem to be inadequate for informed participation in the nation's civic affairs (Mullis and Jenkins, 1988, p. 6).

> Next June, three and a half million 13-year-olds will finish the 7th or 8th grade in the United States. One and a half million of them–almost half–will still not understand basic information from the life and physical sciences... This means that each year, unless things change, one and a half million young Americans will leave their middle school experiences unprepared for secondary school science courses (Anrig and Lapointe, 1989, p. 7).

> A key point...is that an over reliance on multiple-choice testing not only emphasizes simple recall of facts and/or recognition of textbook experiments, but militates against the less predictable hands-on approach... This claim finds support in data from the NAEP Report Card showing that the percentages of students in high school (age 17) who were at or above proficiency levels on higher order thinking skills in science were shockingly low. Only 7% can infer relationships and draw conclusions using detailed knowledge and 41% have some detailed scientific knowledge and can evaluate the appropriateness of scientific procedures (Baron et al. 1989, p. 2).

---

[1]Daniel Koretz (1989) appears to be one of the first researchers to question the validity of some of the interpretations derived from the results reported for the NAEP proficiency scales. In a paper entitled "NAEP's Scales: How Useful Are They?", presented at the 1989 ECS/CDE annual conference, he discusses some of the issues considered in the present paper. Dan graciously shared his notes for that presentation with me.

The primary purpose of this paper is to consider the validity of statements such as these. It is contended that, although these generalizations may be true, the current NAEP results provide little, if any, support for them. It is also contended that the purported "criterion-referenced characteristic" of the NAEP scales actually invites such inappropriate inferences. Presumably, the procedures used to develop the NAEP scales support generalizations about what students can and cannot do. I will argue, however, that the NAEP scales, as constructed, do not yield meaningful criterion-referenced interpretations.

The major focus of these remarks is on the reasonableness (or unreasonableness) of the interpretations derived from the very complex NAEP scaling procedures and not on the details of these complex procedures. In the next section, I will describe the major goal of the NAEP scales and summarize briefly the procedures by which NAEP attempts to attach meaning to the scales. The NAEP science scale is then considered in some detail, followed by a few observations about three other NAEP scales.

GOAL OF THE NAEP PROFICIENCY SCALES

The NAEP scales have been developed primarily to describe student achievement in specific content areas, such as science, from grade 3 (age 9) through grade 11 (age 17). Thus, in some general sense, the purpose of the NAEP scales is highly similar to the purpose of the developmental score scales derived for multilevel standardized achievement tests (e.g., the scales of the *Comprehensive Tests of Basic Skills*, CTB/McGraw Hill, 1989). Obviously, any attempt to construct such a developmental scale requires a considerable number of arbitrary decisions. Standardized achievement test authors, for example, must make judgments about appropriate content and the sequence of that content for a particular curriculum area. Other judgments must be made about the proportion of items at various cognitive processing levels, such as recall vs. analysis, for each level of the test. The end result of these decisions provides the operational definition of the achievement construct of interest for these authors—the authors' "test specifications." These same types of judgments and decisions are also an integral part of the NAEP developmental process. The specific process used by NAEP (sometimes

called a consensus process) to make these arbitrary decisions and the "test specifications" resulting from them are described in various NAEP publications.[2]

Although the substantive differences between the procedures followed by NAEP and standardized achievement test publishers in developing test specifications are relatively minor, differences in the proposed interpretation of the scores derived from the resultant scales are fairly dramatic. As is well known, the interpretation of the numerical values associated with standardized achievement tests is based primarily on norm-referencing. Thus, on the basis of a particular score it might be noted that this examinee's performance is similar to that of the typical grade 7 student. Or, it might be observed that mathematics problem solving is a relatively weak area for this examinee. Or, it might be concluded that the "science growth" shown by this examinee since the last testing was typical for students at that age or grade.[3] Of course, the interpretation of a student's performance might be enhanced by looking at illustrative item clusters associated with the content and/or process categories in the test specifications. And, at this level of analysis, both criterion-referenced and norm-referenced interpretations might prove somewhat useful.

The numerical values for the NAEP scales, on the other hand, are intended to be interpreted almost exclusively from a criterion-referenced perspective. Most NAEP Report Cards contain a statement similar to the following, which was taken from the *Science Report Card*:

> One of NAEP's major goals has always been to identify what students know and can do and stimulate debate about whether those levels of performance are satisfactory. An additional benefit of IRT methodology is that it provides for a criterion-referenced interpretation of levels on a continuum of proficiency. (Mullis and Jenkins, 1988, p. 141)

Thus, a major goal of the NAEP scale is to identify what students can do, clearly a worthwhile and desirable goal.

---

[2]E.g., the process used to develop the 1986 science assessment and the details about the test specifications are described in: *Science Objectives: 1985-86 Assessment* (ETS, 1986).

[3]Such interpretations, of course, are inextricably linked to a particular set of test specifications and the implementation of these specifications.

Developing scales that permit criterion-referenced interpretations for examinee scores was a measurement goal prior to Glaser's famous 1963 *American Psychologist* article on this topic; it remains a goal for many assessment purposes today. NAEP claims to accomplish this goal through its use of IRT methodology, since one outcome of this methodology is the common scaling of items and ability (Reckase, 1989). This link between individual items and the ability dimension is used to describe the nature of the NAEP proficiency scales. How this is accomplished is briefly described in the statement below:[4]

> Although the proficiency scale ranges from 0 to 500, few students performed at the ends of the continuum. Thus, levels chosen for describing results in the report are 150, 200, 250, 300, and 350. Each level is defined by describing the types of science questions that most students attaining that proficiency level would be able to perform successfully; each is exemplified by typical benchmark items...
>
> In the scale-anchoring process, NAEP identified sets of items from the 1986 assessment that were good discriminators between proficiency levels. The guideline used to select such items was that students at any given level would have at least a 65 to 80 percent (but often higher) probability of success with these science questions, while the students at the next lower level would have a much lower probability of success [less than 50 percent] using the criterion that the difference in probabilities exceeds 30 percent between adjacent levels. Science specialists examined these empirically selected item sets and used their professional judgment to characterize each proficiency level (Mullis and Jenkins, 1988, pp. 141-142).[5]

Note that the only basis for classifying an item at a proficiency level is the estimated probability of success of examinees at different proficiency levels. Most importantly, note that the subject matter and the cognitive skill addressed by the item are not directly considered in this classification. These two factors are involved, however, when the descriptions for each level are derived. The implications of this procedure for the criterion-referenced interpretations of the NAEP score values are discussed next within the context of the NAEP Science Proficiency Scale.

---

[4]This statement is from the *Science Report Card*. Similar statements are given in the report cards for other assessment domains.

[5]More details related to the anchoring process are provided in Mullis (1990).

THE NAEP SCIENCE PROFICIENCY SCALE

Figure 1 presents the primary science scale used for the reporting of NAEP science data. [Figure 1 and all of the NAEP science material reproduced here are taken from either the NAEP *Science Report Card* (Mullis and Jenkins, 1988) or *Science Objectives: 1985-86* (ETS, 1986).] It should be noted that this scale is really a composite scale based on five subscales.[6] (Only three subscales are used at age 9.) Presumably, students who took the NAEP science items received a score on each of these subtests and these were weighted in a specific way to obtain an estimate of the science proficiency level for each student. Then, the procedure outlined in the preceding section was used to provide the description of "what examinees at each of the five levels of proficiency really can do."

Before considering some of the interpretation problems associated with the Science Proficiency Scale, it seems appropriate to review some of Nitko's (1984) observations about criterion-referenced measurement. Nitko (p. 13) classified the domains used by test developers into four types: (1) well-defined and ordered domains; (2) well-defined but unordered domains; (3) ill-defined domains, and (4) undefined domains. He noted that ill-defined domains have "poorly articulated behavioral objectives" and that the domain tends to be defined "only in terms of the particular items in the test" (p. 13). He also observed:

> Ill-defined domains and undefined domains cannot form the basis for building a criterion-referenced test, and so tests developed from these two categories cannot be called criterion-referenced under the broad definition adopted here, even though some test developers may claim otherwise (pp. 12 and 14).

A final observation from Nitko seems especially pertinent to the NAEP scales because it concerns the use of latent trait [item response] methodology to develop a criterion-referenced test:

---

[6]Some insights into what this composite scale might represent can be obtained from the two publications noted. The results and the interpretations of these results that usually appear in popular publications are based on this composite scale.

Level 150—Knows Everyday Science Facts

Students at this level know some general scientific facts of the type that could be learned from everyday experiences. They can read simple graphs, match the distinguishing characteristics of animals, and predict the operation of familiar apparatus that work according to mechanical principles.

Level 200—Understands Simple Scientific Principles

Students at this level are developing some understanding of simple scientific principles, particularly in the Life Sciences. For example, they exhibit some rudimentary knowledge of the structure and function of plants and animals.

Level 250—Applies Basic Scientific Information

Students at this level can interpret data from simple tables and make inferences about the outcomes of experimental procedures. They exhibit knowledge and understanding of the Life Sciences, including a familiarity with some aspects of animal behavior and of ecological relationships. These students also demonstrate some knowledge of basic information from the Physical Sciences.

Level 300—Analyzes Scientific Procedures and Data

Students at this level can evaluate the appropriateness of the design of an experiment. They have more detailed scientific knowledge, and the skill to apply their knowledge in interpreting information from text and graphs. These students also exhibit a growing understanding of principles from the Physical Sciences.

Level 350—Integrates Specialized Scientific Information

Students at this level can infer relationships and draw conclusions using detailed scientific knowledge from the Physical Sciences, particularly Chemistry. They also can apply basic principles of genetics and interpret the societal implications of research in this field.

Figure 1. NAEP Science Proficiency Levels (Mullis and Jenkins, 1988, p. 58)

> Sometimes the domain of behavior represents a single dimension or factor, called a *latent trait*, which is hypothesized to underlie the performance of specific behaviors. If so, it might be possible to scale the tasks in the domain along this dimension using one of the various latent trait models....
>
> Note that it is not latent trait analysis in general that is criterion-referenced. Rather, latent trait analysis might be used to help order the tasks comprising a domain so that an examinee's test score may be referenced to the domain in a way that reveals the specific behaviors of which the examinee is capable. This does not flow automatically from an application of latent trait measurement theory, but *requires that the domain of tasks be unidimensional and that the items (tasks) be orderable on a number line representing this dimension* [Italics added]. The resultant test scores must be capable of being interpreted in terms of the specific behaviors in the domain an examinee is likely to be able to perform (p. 17).

I would contend that the NAEP science domain is an ill-defined domain at best. It consists of learning outcomes from a variety of content areas and at a variety of process levels. For example, the assessment framework presented in the *Science Objectives* booklet identifies 6 content areas, 3 cognitive levels and 4 context situations, a total of 72 unique combinations for developing items. Furthermore, it is highly doubtful that learning within this domain will proceed through neat sequential stages along some ordered dimension. In addition, the number of test items that could be developed for this domain is virtually unlimited.[7] Finally, test developers face the enormous problems created by the interaction of an examinee's past experiences and the content of the item. This interaction becomes particularly critical when developing a scale for use with students at several age/grade levels and definitely impacts the nature of what is being assessed. Consider, for example, the following statement from the *Science Objectives* booklet (ETS, 1986):

> In the cognition dimension [3 categories], the committee assigned the same percentages to all three ages, with the understanding that what is expected of 9-year-olds in the "integrates" category (which calls for multi-step processes and higher-order skills) will be significantly different from what is expected of older students. Since it is not possible to know precisely what knowledge and processes any individual student will apply to a given exercise, the classification of the exercises as "knows" or "uses" or even "integrates" was based on informed judgment about the factual knowledge and cognitive processes that an average student in the target populations is most likely to apply to obtain the correct answer. For example, for one student, answering a particular question may simply involve knowing the factual information needed. Another who does not have that specific information may arrive at the correct answer by a mental

---

[7]These observations could be made about each of the NAEP domains considered in this paper. Gronlund (1973) and Gronlund and Linn (1990) make similar observations about the measurement of student achievement with respect to "developmental objectives."

> process that involves different but related information. Everyone who classified exercises needed to consider the age of the intended population as well as its probable academic experience. An exercise can be in the "knows" category for older students and in the "integrates" category for younger ones, because what older students may know as a fact younger students probably can arrive at only by multi-step reasoning (pp. 10-11).

The following question, taken from page 31 of the objectives booklet, illustrates this situation:

> In which set of living things below do all four things get their food in a similar way?

After this item is presented, the authors write:

> This exercise for 9-year-olds is classified as Life Science–Scientific–Integrates. (It is an example of a question that would be classified as integrates for 9-year-olds but not older students.) (p. 31)

Doesn't this comment acknowledge the absence of a well-ordered domain along a single dimension? How is such a situation resolved within the NAEP scaling procedure which derives a single proficiency scale to be used to define what students from age 9 through age 17 can do? Or, more generally, starting with a framework for the science assessment that has three major dimensions (cognition, context, and content), how is a single scale developed to describe what students from ages 9 through 17 can do? The descriptions given in Figure 1 provide some help in answering these questions. Note, for example, that Level 350 is characterized by integration (cognitive), Chemistry/Genetics (content), and societal implications (context). The descriptions of all five levels exhibit a mix of the content and cognition dimensions and, at times, the context dimension. In fact, the cognition dimension seems to have increased from the three categories identified during the planning phase (knows, uses, integrates) to five during the scaling phase (knows, understands, applies, analyzes, integrates), and the content dimension has decreased from six categories (life science, physics, chemistry, earth and space sciences, history of science, and nature of science) to basically four (everyday experiences [content or context?], life sciences, physical sciences, chemistry). A very abbreviated description of this scale would be as follows:

Level	Process	Content
150	Knows	Everyday experiences
200	Understands	Life Sciences
250	Applies	Life Sciences/Physical Science
300	Analyzes	Physical Science
350	Integrates[8]	Physical Science, particularly Chemistry

Given the method used to develop these scales, it appears that it would have been impossible for the illustrative item above to be classified as a Level 350 item, since it is based on content in the life sciences. Were any items involving life sciences classified as Level 350 items? Furthermore, given the proficiency scale descriptions it might not be surprising that the following item is used to "anchor" Level 350:

> Elements with chemical characteristics most similar to those of sodium are listed in what part of the periodic table?
>
> ANS. Above and below sodium in the same column.

For the Russian chemist Mendeleyev, derivation of the periodic table was undoubtedly an astounding feat of integration. However, it is difficult to conceive of this item as measuring a high school student's ability to integrate specialized scientific information. Rather, it appears simply to require knowledge of how the periodic table is organized.[9]

Actually, within each of the defined proficiency levels there are items that seem to measure primarily recall of facts, concepts, principles, or generalizations. Consider, for example, the following items:

---

[8]It is interesting to note that the definition of integration provided on pages 9 and 10 of the *Science Objectives* booklet doesn't include specific content areas such as physical science, genetics, etc.
[9]Koretz (1989) also makes this observation.

Level 200: What is the main function of the heart?

ANS. To pump the blood to all parts of the body

Level 250: In an ordinary light bulb with a screw type base, which is the part that glows to produce the light?

ANS. A special thin wire at the center of the light bulb

Level 300: Which of the following is the best indication of an approaching storm?

ANS. A decrease in barometric pressure[10]

This mixing of dimensions would not be so critical if NAEP did not claim that this scale permits us to describe specifically what individuals can do at several different age levels. Is it surprising, for example, given the nature of this scale, that only 0.1% of 9-year-olds and 0.2% of 13 year-olds are at Level 350 and above. If items like the periodic table item must be answered to reach Level 350 and if items like the "living things" item do not "count" as Level 350 items (integrates), are these 0.1% and 0.2% values unexpected? How relevant is such information? Surely, some significant proportion of our 9-year-olds and 13-year-olds can, to a reasonable degree, "infer relationships and draw conclusions" when using scientific knowledge that it is sensible to expect them to know. Wouldn't information related to this more restricted science domain be of more use to educators and policymakers?

---

[10]As the earlier statement from the NAEP *Science Objectives* booklet implied, when judging the level of cognitive processing required to answer such items, an evaluator makes certain assumptions about the background experiences of the examinee. For example, because reasonably novel tasks must be considered by examinees before we can conclude that they can apply, analyze, or integrate, evaluators must make assumptions about the prior experiences of the examinees with respect to the given tasks. In my evaluation of the Level 300 item, for example, I assume that students at most levels of science instruction have learned that falling barometric pressure indicates approaching storms. Thus, I assume that for most students this item is merely measuring the recall of that concept. No novel task is being faced by the majority of examinees. The NAEP science educators who were "anchoring" this scale obviously concluded that this item required students to analyze scientific procedures and data.

To belabor this point, the question about the structure of the periodic table probably also represents the recall of information learned by the majority of examinees who have taken a general chemistry course. It is difficult to conceive of this question as measuring the integration of scientific information without assuming that the examinees have a fair amount of background knowledge with regard to the chemical properties of elements, basic knowledge of the periodic table, etc. And, if these examinees have this type of background information, it is difficult to believe they haven't considered the structure of the periodic table explicitly. Furthermore, for those examinees without some exposure to material in a formal chemistry course, it would appear that answering this item correctly would be almost impossible.

Failing to recognize this mixing of dimensions can lead reporters, legislators, and even professional educators to drawing very questionable conclusions from the NAEP results. For example, consider the statement by Baron et al. given on page 1. Their estimate that only 7% of 17-year-olds can infer relationships and draw conclusions using detailed knowledge is probably based on data reported in Table 2.1 of the NAEP *Science Report Card* (p. 39). According to data presented in that table, only 7.5% of the 17-year-olds in 1986 had science proficiency levels of 350 or above. However, given the periodic table example above, given the other released items for Level 350, and given the description of the anchor points, it does not seem too unreasonable to conclude that the major reason most of the 17-year-olds cannot integrate specialized scientific information is because they have not had a formal chemistry course. In essence, they do not have the subject matter knowledge required to do the integration. More appropriately, perhaps, in this case, they do not have the subject matter knowledge to answer knowledge questions about chemistry.

The other two statements quoted on page 1 also seem to accept the processing level indicator as defining the level without regard to other dimensions and to attach a great deal of surplus meaning to these indicators. The ill-defined nature of these subdomains never appears to be of concern. The Anrig and Lapointe conclusion is also derived from data reported in Table 2.1 of *The Science Report Card*. In this table, it is estimated that only 53.4% of thirteen-year-olds in 1986 have reached a proficiency level of 250. Evidently, Anrig and Lapointe believe that the ability to apply basic scientific information as defined by the NAEP scales is necessary to be prepared for secondary school science courses. Ignoring obvious issues related to the ability of secondary teachers to adapt instruction and to the definition of success in these science courses, what validity evidence do Anrig and Lapointe have for this inference? Why do they believe this is the level of proficiency necessary for success? What is it in the description of Proficiency Level 250 that permits such a powerful generalization? What criterion-related validity evidence exists to support their conclusion? This might be a very reasonable generalization, but it clearly cannot be supported by the so-called criterion-referenced nature of the NAEP scales or by the results reported. Surely, such important decisions about educational programs

must be made on better evidence than the percent of students estimated to be at or above a certain point on scales as ambiguous as these.

A similar, but in my opinion even more outlandish, interpretation occurs in the statement by Mullis and Jenkins. Again, this statement appears to be based on data in Table 2.1 of the *Science Report Card*, which indicates that only 41.4% of 17-year-olds in 1986 had a Science Proficiency Level of 300 or above. What empirical data support the statement that people with Science Proficiency scores below 300 do not have the necessary skills for informed participation in public affairs. Or, for that matter, that those above 300 have such skills?[11]

The three examples above also ignore the continuous nature of the proficiency scales. Are students with a proficiency score of 240 prepared for secondary science courses? What about students with a 225? What can students with a proficiency score of 250 do that students with a proficiency score of 225 (1/2 standard deviation lower) can not do? Likewise, can students with a proficiency level of 325 integrate specialized scientific information? If so, what percent of 17-year-olds can now "infer relationships and draw conclusions?" Is this percent reasonable given that some formal instruction in chemistry is probably required to reach this level?

Given the obvious concerns about making valid criterion-referenced interpretations from a proficiency scale developed for such an ill-defined domain, it is interesting to speculate how this scale might be used if the NAEP assessment procedures were implemented so that school district and, perhaps, even individual scores were reported. For example, suppose that in addition to the data gathered at ages 9, 13, and 17, data were also collected at ages 10, 11, 12, 14, 15, and 16. Further suppose

[11]Another confounding factor in these types of interpretations relates to the effort put forth by the examinees. Students taking the NAEP exercises receive no information about their performance. Nor is school level performance reported. Thus, these interpretations assume that examinees are making a sincere effort even though they have no stake in the assessment. This altruistic assumption may not be tenable for a significant proportion of examinees, particularly the 17-year-old examinees.

that the mean science proficiency scores for these various age groups were as follows:

Age	(Grade)	Mean*
9	(3)	224
10	(4)	230
11	(5)	237
12	(6)	244
13	(7)	251
14	(8)	260
15	(9)	269
16	(10)	279
17	(11)	289

*The mean values for ages 9, 13, and 17 are given on p. 24 of the *Science Report Card*. The other means were estimated by interpolation.

Notice that only *one* of the five anchor level points (250) is included within this range of mean values! How would values of 235, 236, etc. be interpreted by schools and students? A guess: We would have the NAEP Age-Equivalent or Grade-Equivalent scales. Thus, for example, a score of 235 would be interpreted as a performance level similar to the average fifth grade student. And, as is true with standardized achievement tests, some additional information about the types of items such a student answered correctly and incorrectly would be provided to help enhance this interpretation. I would predict that people would very quickly stop using the criterion-referenced interpretations supposedly available from the five anchor level descriptions.

MATHEMATICS PROFICIENCY SCALE

Figure 2 presents the primary mathematics scale used for reporting NAEP math results. As was true with the primary science scale, this scale is actually a composite scale based on varying numbers of subtests at each age/grade level. (See Dossey et al. 1988 for more details.)

Level 150—Simple Arithmetic Facts

Learners at this level know some basic addition and subtraction facts, and most can add two-digit numbers without regrouping. They recognize simple situations in which addition and subtraction apply. They also are developing rudimentary classification skills.

Level 200—Beginning Skills and Understanding

Learners at this level have considerable understanding of two-digit numbers. They can add two-digit numbers, but are still developing an ability to regroup in subtraction. They know some basic multiplication and division facts, recognize relations among coins, can read information from charts and graphs, and use simple measurement instruments. They are developing some reasoning skills.

Level 250—Basic Operations and Beginning Problem Solving

Learners at this level have an initial understanding of the four basic operations. They are able to apply whole number addition and subtraction skills to one-step word problems and money situations. In multiplication, they can find the product of a two-digit and a one-digit number. They can also compare information from graphs and charts, and are developing an ability to analyze simple logical relations.

Level 300—Moderately Complex Procedures and Reasoning

Learners at this level are developing an understanding of number systems. They can compute with decimals, simple fractions, and commonly encountered percents. They can identify geometric figures, measure lengths and angles, and calculate areas of rectangles. These students are also able to interpret simple inequalities, evaluate formulas, and solve simple linear equations. They can find averages, make decisions on information drawn from graphs, and use logical reasoning to solve problems. They are developing the skills to operate with signed numbers, exponents, and square roots.

Level 350—Multi-step Problem Solving and Algebra

Learners at this level can apply a range of reasoning skills to solve multi-step problems. They can solve routine problems involving fractions and percents, recognize properties of basic geometric figures, and work with exponents and square roots. They can solve a variety of two-step problems using variables, identify equivalent algebraic expressions, and solve linear equations and inequalities. They are developing an understanding of functions and coordinate systems.

Figure 2. NAEP Mathematics Proficiency Levels (Dossey et al. 1988, p. 31)

The math proficiency scale exhibits many of the same problems noted with the science proficiency scale and it does so for exactly the same reasons. That we are dealing with a very ill-defined domain seems clear just by considering the descriptions of the various levels. The assessment framework given in *Math Objectives: 1985-86 Assessment* (ETS, 1986) provides additional support for this conclusion—seven content areas and five process categories are identified in this booklet (p. 8). Thus, again, the multidimensional nature of the domain will probably create interpretation problems if the proficiency scale values are used to make criterion-referenced interpretations across age/grade levels.

As an example of the mixing of the content and process dimensions, note that the label for Level 250 is, in fact, a combination of content (Basic Operations) and process (Beginning Problem Solving). The complete description provided for this level provides additional evidence of this mix.

As was true with the science scale, the only data used to support the development of this particular scale are the estimated probabilities of success on the benchmark items for students at various proficiency levels. It is interesting to speculate what might have happened if mathematics educators had been given the items without the a priori classifications based on probabilities of success and asked to order them along some continuum of mathematics development. For example, suppose the seven items shown in Figure 3 had been included in such an experiment. Would most math educators have clustered the first five items together at the same point along the proficiency continuum? And, if they had, would they have labeled them as representing "Moderately Complex Procedures and Reasoning"? These are 5 of the 6 released items that help define Level 300 for the NAEP scale![12] Would these math educators have labeled item #6 as an item that represents "Multi-Step Problem Solving and Algebra"? And would they have been more likely to cluster item #2 (Level 300) with item #7 (Level 350) than with the other Level 300 items shown in Figure 3? These few examples should illustrate that expecting this scale to describe what students can do in any meaningful way is not very realistic.

---

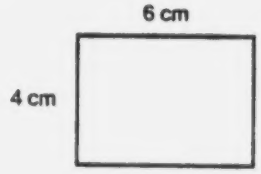
[12]The other released item is a fairly simple graph-reading item requiring some extrapolation to arrive at the correct answer.

1. Which of the following is true about 87% of 10?

   ANS. It is less than 10

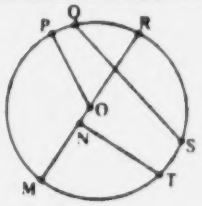
2. If $7X + 4 = 5X + 8$, then $X =$

   ANS. 2

3. 

   What is the area of this rectangle?

   ANS. 24 square cm



4. Which of the following is radius of the circle?

   ANS. $\overline{OP}$

5. Which points are the end points of an arc?

   ANS. Q, S

---

6. $\sqrt{17}$ is between which of the following pairs of numbers?

   ANS. 4 and 5

7. Which of the following are equivalent equations?

   ANS. $Y - 3 = 7$ and $Y + 5 = 15$

Figure 3 NAEP Mathematics Items (Dossey et al., 1988)

My contention that the nature of the NAEP scales actually invites misinterpretation is illustrated by the following statement:

> Although the knowledge and problem-solving skills required to answer items at Level 300 are too advanced for 9-year-olds, it is troubling that more 13- and 17-year-olds have not attained this level of performance. Given that students are exposed to many of these topics in middle and junior high school, one would expect to see a higher percentage of students at age 13 and particularly at age 17 demonstrating success at this level of proficiency. *The findings seem to lend support to recent calls for more challenging curriculum in the middle and upper grades* [Italics added] (Dossey et al. 1988, p. 40).

The empirical data to support this conclusion are presumably given in Table 2.1 of the *Mathematics Report Card*. In that table it is reported that only 16% of the 13-year-olds and 51% of the 17-year-olds have reached Level 300 or above in 1986. It appears that the interpretation of this finding as supporting a call "for a more challenging curriculum in the middle and upper grades (a call that many people would probably make without the NAEP data) is based on the label given to Level 300: Moderately Complex Procedures and Reasoning. However, the released items hardly seem to measure reasoning and complex procedures. (See Figure 3.) In fact, one could argue that three of these items represent fairly basic geometric concepts and that none of the six really requires much reasoning. Thus, a valid conclusion to be derived from these data would perhaps relate more to the knowledge of mathematical concepts than to reasoning. And unless the new and more challenging curriculum does a better job of helping students learn fundamentals of mathematics, we have little hope for producing the type of mathematically literate graduates needed for our increasingly technological society.

One of the most dramatic misinterpretations of NAEP results is provided by Shanker (1990). In this article, which was concerned with the restructuring of our public schools, Shanker uses NAEP results as the primary evidence for answering the question, "How bad are things?" He uses data from a variety of NAEP assessments; however, only the mathematics example is given below. The nature of the erroneous interpretation generalizes to the other NAEP assessment areas. Shanker notes:

> The assessment results for 17-year-olds who are still in school are particularly dismaying. Most of the 25% of high school students who drop out are gone. The 17-year-olds who are there to be tested are our successful students, the ones who are about to march down the aisle and get diplomas. Yet the findings of the NAEP indicate that few of these students are ready to do real college-level work or to handle a good job. For

> example, only 6% could solve the following multi-step math problem: "Christine borrowed $850 for one year from the Friendly Loan Company. If she paid 12% simple interest on the loan, what was the total amount she repaid?" (p. 36)

Shanker used the results reported in the *Mathematics Report Card* to arrive at the 6% value. The item quoted by Shanker is a "Level 350" item, and Table 2.1 in the report card indicates that only 6.4% of the 17-year-olds are at proficiency Level 350 and above. Thus, Shanker concludes that only 6% of the 17-year-olds in our schools can do this problem. However, this is unequivocally an erroneous interpretation of the data and probably drastically misrepresents the situation.[13] It occurs because Shanker interprets the percent at or above a certain proficiency level as the p-value for items at that level. NAEP has estimates of the actual p-values for these items but they are not reported in the *Mathematics Report Card*. The estimated p-value for this item will certainly be greater than .06, and if calculators had been permitted, an even greater p-value would probably have been obtained.

I would contend that the purported characteristics of the NAEP scale have led to Shanker's misinterpretation. If we say these scales tell us what students can and can not do, then taking the percent of students at Level 350 and above and interpreting this as the percent of students who can do Level 350 items is not an unreasonable generalization to make.

---

[13]This misinterpretation was first brought to my attention by H.D. Hoover on February 28, 1990 at a measurement conference sponsored by Riverside Publishing Company in Atlantic City. Subsequently, Robert Linn identified the same misinterpretation in his remarks as a discussant for a NAEP symposium at the 1990 AERA annual meeting.

OTHER NAEP SCALES

A few observations about the criterion-referenced interpretations derived from the NAEP reading and U.S. history scales are given in this section. These two scales differ from the science and mathematics scales in that they are not composite scales derived from a set of subtest scales. Thus, these domains are not as ill-defined as the domains for science and mathematics; however, they are not well-defined domains, and the problems associated with making valid criterion-referenced interpretations of the scale scores still remain. For example, the U.S. history scale consists of the following four levels:

Level 200: Knows Simple Historical Facts

Level 250: Knows Beginning Historical Information and Has Rudimentary Interpretative Skills

Level 300: Understands Basic Historical Terms and Relationships

Level 350: Interprets Historical Information and Ideas.

What underlying well-ordered, developmental construct permits us to differentiate between "Knows Simple Historical Facts" (Level 200) and "Knows Beginning Historical Information" (Level 250)? How does "Understanding Basic Historical Terms" (Level 300) differ from "Knows Beginning Historical Information" (Level 250)? Given these brief descriptions, at what level would each of the following items from the *U.S. History Report Card* (Hammack et al. 1990) be classified?

1. Soldiers fighting for the South during the Civil War were called

   ANS. Confederates

2. Which of the following is the most recent invention?

   ANS. Space shuttle[14]

---

[14]Throughout this paper, I have presented only the answers for the illustrative NAEP items–the distractors were not provided. It is well known that the particular set of alternatives for an item can impact various item characteristics, including, in some instances, the cognitive process being assessed. For example, one of the alternatives for this particular item was "Covered Wagon." Would this item still be an anchor item for Level 250 if a more plausible alternative had been used?

3. The most important reason the United Nations was organized after the Second World War was to help countries

   ANS. keep peace

Does it seem reasonable that item #1 is a Level 300 anchor item, whereas item #3 is a Level 200 anchor item? (Item #2 is a Level 250 anchor.) What developmental continuum is being defined by such differentiations? Given these illustrative items, one at each of three different levels, it is not surprising that many of the interpretations derived from the 1988 NAEP U.S. History assessment are based primarily on specific data about individual items. (See Hammack et al. 1990, pp. 7-10)

A final observation concerning the U.S. History Proficiency Scale relates to the Level 350 anchor items. The seven released items defining this level appear to be measuring examinees' familiarity with facts, concepts, or generalizations that are probably encountered relatively late in their academic career rather than examinees' ability to interpret historical information or ideas. Consider for example, the following three items:

4. Jane Addams founded Hull House in Chicago in 1889 primarily to

   ANS. improve the community and civic life of the urban poor

5. What do Gloria Steinem, Betty Friedan, and Kate Millett have in common?

   ANS. They all have written books and articles in support of the women's movement.

6. Formal diplomatic ties between the United States and the People's Republic of China were established during the presidency of

   ANS. Richard M. Nixon

Is it plausible to claim that anyone can describe what students can do in the way of interpreting historical information and ideas on the basis of such items? I think not.

The area of reading assessment appears to provide a good opportunity for developing scales that permit users to draw valid criterion-referenced interpretations. Certainly in some ways this domain is less ill-defined than the others. However, NAEP has not been much more successful in this area than in other areas discussed above. In fact, the Reading Proficiency Scale reflects some of the same problems noted above. Specifically, the mix of the process and content dimensions again creates problems. Mullis and Jenkins (1990) note that "the interaction of three factors affects students' reading

proficiency: the complexity of the material [content], their familiarity with the subject matter [content], and the kinds of questions asked [reading skill or process]" (p. 22).

Figure 4 contains three items that would appear to be measuring literal comprehension. (The answers and the relevant text material are also given in Figure 4.) Note, however, that these items were used to help "anchor" three different levels of proficiency and that this labeling must have occurred primarily because of the nature of the reading material and not the process skill being measured.

Level 200 Basic . . . *Performance at this level suggests the ability to understand specific or sequentially related information.*

Q. What is quicksand?

ANS. Soupy sand you can't stand on

Text: Quicksand often looks like sand. But it is really soupy sand with so much water that you can't stand on it.

Level 250 Intermediate . . . *Performance at this level suggests the ability to search for specific information, interrelate ideas, and make generalizations.*

Q. Who invented the game of basketball?

ANS. A Massachusetts teacher

Text: When Dr. James A. Naismith, a teacher at the international YMCA Training School in Springfield, Massachusetts, first invented the game...

Level 300 Adept . . . *Performance at this level suggests the ability to find, understand, summarize, and explain relatively complicated information.*

Q. In what year did the first United States congresswoman take office?

ANS. 1917

Text: In 1917 New York followed the example of the western states. In that same year Jeannette Rankin of the state of Montana took office as the first United States congresswoman.

Figure 4. NAEP Reading Items (Mullis and Jenkins, 1990)

The key phrase for Level 350 (the highest level reported) is: "Performance at this level suggests the ability to synthesize and learn from specialized materials" (Mullis and Jenkins, 1990, p.23). According to the NAEP results only 0.2% of age 13 students are at or above Level 350 (p. 33). Thus, according to these results very, very few seventh grade students can synthesize and learn from "specialized materials". No independent readers here! This conclusion is ridiculous, as I think most seventh grade teachers would agree. The inference is drawn only because of the decision to try to impose a unidimensional proficiency scale on a multidimensional domain.[15]

On the basis of the 1984 NAEP reading assessment, Applebee et al. (1987) made the following statement based on the report that only about 21% of young adults had reached Level 350: "Such tasks require the ability to reason effectively about what is read—and few people were able to do that" (p. 16). Whether or not Level 350 tasks truly require students to reason or just force them to deal with larger chunks of material in some literal way may be debated. However, it is probably true that the Applebee et al. statement will be taken literally. Consider, for example, the statement by David Kearns, Chairman and Chief Executive Officer of Xerox Corporation, in the Foreword to the Applebee et al. booklet:

> According to the National Assessment of Educational Progress, only a small percentage of the young people sampled in its recent studies can reason effectively about what they read and write. That means that the majority don't have the critical thinking skills we need in an economy like ours . . . (p. 3).

Kearns seems to have focused on the Applebee et al. statement concerning Level 350 tasks and the fact that only about 5% of grade 11 students are at or above this level and 0% of both grade 8 and grade 4 students reached this level (Applebee et al. p. 15). Hence, once again we have an example of the acceptance of the superficial criterion-referenced interpretations accompanying the NAEP proficiency scales as really describing what students can do and of the attachment of considerable surplus meaning to these interpretations. It is difficult to find fault with Kearns' interpretation, however, given the equally invalid interpretations made in the NAEP publications.

---

[15]Other factors (dimensions) that appear to create interpretation problems in this area are labeled context effects. At the 1990 AERA symposium, "An Update on the NAEP 1986 Reading Anomaly," context effects (e.g., item order and context) were suggested as the main cause of the famous NAEP reading anomaly.

CONCLUDING STATEMENT

For any content domain, the construction of a multilevel achievement test and the associated developmental score scale represents an enormous challenge. When the domain is ill-defined, the challenge becomes even greater. Inevitably, when different efforts are undertaken to meet this challenge, different operational definitions of what is to be measured and how it is to be measured will occur. If this were not true, concerns about which of several multilevel standardized achievement tests has the "best curriculum alignment" would not be such a critical issue in the selection of these tests. The relatively recent debate concerning the "best scale" to use when measuring the educational development of elementary school students also illustrates that a number of fundamental issues related to the scaling of such achievement domains have not been adequately resolved (Burket, 1984; Hoover, 1984a; Hoover, 1984b; Yen, 1986; Hoover, 1988; Phillips and Clarizio, 1988; Yen 1988).

No one challenges the idea that it would be desirable to have educational measurements which accurately describe what examinees can and can not do (Forsyth, 1976). Teachers have pleaded for such measures for decades. Providing such measures was and is one of the goals of the NAEP proficiency scales. The purpose of this paper was not to criticize this goal.[16] Rather, the major purpose was to show that NAEP, despite its claims, has not achieved this goal to any reasonable extent. Indeed, it was proposed that such a goal is unattainable given the ill-defined NAEP domains and our present knowledge base. Even the use of IRT methodology cannot overcome these obstacles. In fact, the methodology may exacerbate problems of criterion-referenced interpretations for such domains. A secondary purpose of the paper was to illustrate some of the inappropriate interpretations of the NAEP data that occur, in my opinion, primarily because NAEP has lead people to believe that this goal has been attained.

---

[16]It should be noted that this paper also was not concerned with validity issues related to other uses of the NAEP scales–for example, the usefulness of the NAEP scales for describing change in achievement over time. Likewise, no criticism of NAEP's efforts to develop test specifications for a given domain was intended. In fact, ignoring cost effectiveness issues, NAEP's consensus approach seems very reasonable.

Do the NAEP proficiency scales yield valid criterion-referenced interpretations? As many measurement experts have observed, validity is not an all or none concept. Nonetheless, given the observations presented above, I would answer, without reservations, "no."

References

Anrig, G. R. and Lapointe, A. F. (1989). What we know about what students don't know. *Educational Leadership, 47,* 4-9.

Applebee, A. N., Langer, J. A., and Mullis, I. V. S. (1987). *Learning to be literate in America: Reading, writing, and reasoning.* Educational Testing Service, Princeton, NJ.

Baron, J. B., Forgione, P. D., Rindone, D. A., Kruglanski, H., and Davey, B. (1989). *Toward a new generation of student outcome measures: Connecticut's common core of learning assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3,* 15-16.

CTB/McGraw Hill. (1989). *Comprehensive Tests of Basic Skills* (4th ed.). CTB/McGraw Hill, Monterey, CA.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. M., and Chambers, D. L. (1988). *The mathematics report card: Are we measuring up?* Educational Testing Service, Princeton, NJ.

Educational Testing Service. (1985). *The Reading report card: Progress toward excellence in our schools.* Educational Testing Service, Princeton, NJ.

Educational Testing Service. (1986). *Science objectives: 1985-86 assessment.* Educational Testing Service, Princeton, NJ.

Educational Testing Service. (1986). *Math objectives: 1985-86 assessment.* Educational Testing Service, Princeton, NJ.

Forsyth, R. A. (1976). *Describing what Johnny can do.* Iowa Testing Programs, Iowa City, IA.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-21.

Gronlund, N. E. (1973). *Preparing criterion-referenced tests for classroom instruction.* New York, The Macmillan Company.

Gronlund, N. E. and Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). The Macmillan Company, New York.

Hammack, D., Hartoonian, M., Howe, J., Jenkins, L. B., Levstik, L. S., MacDonald, W. B., Mullis, I. V. S., and Owen, E. (1990). *The U.S. history report card.* Educational Testing Service, Princeton, NJ.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3,* 8-14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice, 3,* 16-18.

Hoover, H. D. (1988). Growth expectations for low-achieving students. *Educational Measurement: Issues and Practice, 7,* 21-23.

Koretz, D. (1989). *NAEP's Scales: How useful are they?* Presentation at the annual meeting of the Education Commission of the States and the Colorado Department of Education, Boulder, CO.

Mullis, I. V. S. (1990). *Giving meaning to the IRT mechanizations: The art of anchoring the NAEP scales.* Symposium paper presented at the annual meeting of the American Educational Research Association, Boston.

Mullis, I. V. S. and Jenkins, L. B. (1988). *The science report card: Elements of risk and recovery.* Educational Testing Service, Princeton, NJ.

Mullis, I. V. S. and Jenkins, L. B. (1990). *The reading report card, 1971-1988.* Educational Testing Service, Princeton, NJ.

Nitko, A. J. (1984). Defining "criterion-referenced test." In Berk, R. A. (ed.). *A Guide to criterion-referenced test construction.* The Johns Hopkins University Press, Baltimore, MD.

Phillips, S. E. & Clarizio, H. F. (1988). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice, 7,* 8-15.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8,* 11-15.

Shanker, A. (1990). A proposal for using incentives to restructure our public schools. *Phi Delta Kappa, 71,* 345-357.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325.

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Educational Measurement: Issues and Practice, 7,* 16-17.